PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-250072
(43)Date of publication of application : 17.09.1999

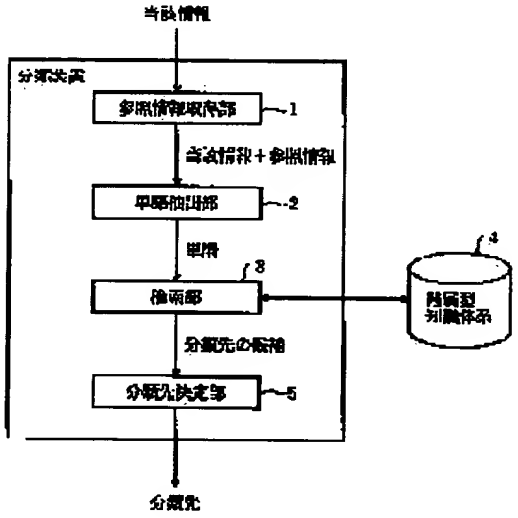
(51)Int.Cl. G06F 17/30
G06F 17/27
// G06F 13/00

(21)Application number : 10-045770
(22)Date of filing : 26.02.1998

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : MURAMOTO TATSUYA
WASHISAKI SEIJI

(54) INFORMATION SORTING METHOD, DEVICE THEREFOR AND STORAGE MEDIUM STORED WITH INFORMATION SORTING PROGRAM

(57)Abstract:
PROBLEM TO BE SOLVED: To solve problems of poor sorting precision caused by the ambiguity of key word extraction and real-time processing due to the need of long learning time by making a single word extracted from a sorting object correspond to an existing hierarchical knowledge system.
SOLUTION: A single word extracting part 2 executes the morpheme analysis of sorting object information acquired by a reference information acquiring part 1 and reference information being information related to it to divide a single word and to give a part of speech to the single word. Among the given parts of speech, a noun and an adjective word are extracted to obtain the occurrence frequency of them to transfer a single word of the highest occurrence frequency to a retrieving part 3. The part 3 retrieves the hierarchical knowledge system 4 by the single word extracted by the part 2 and makes sorting items correspond to the single word to obtain a sorting candidate. A sorting destination deciding part 5 calculates through the use of the occurrence frequency of the single word obtained by the part 2 and the frequency of the sorting item obtained by the part 3 and sorts the value to decide an item becoming a high order to be a sorting destination item.



LEGAL STATUS

[Date of request for examination] 11.01.2001

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－250072

(43)公開日　平成11年(1999)９月17日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/401	３１０Ｄ
	17/27			13/00	３５１Ｇ
// Ｇ０６Ｆ	13/00	３５１		15/38	Ｄ
				15/40	３７０Ａ

審査請求　未請求　請求項の数12　ＯＬ　（全７頁）

(21)出願番号　特願平10－45770

(22)出願日　平成10年(1998)２月26日

(71)出願人　000004226
日本電信電話株式会社
東京都新宿区西新宿三丁目19番２号

(72)発明者　村本　達也
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(72)発明者　鷲崎　誠司
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(74)代理人　弁理士　伊東　忠彦

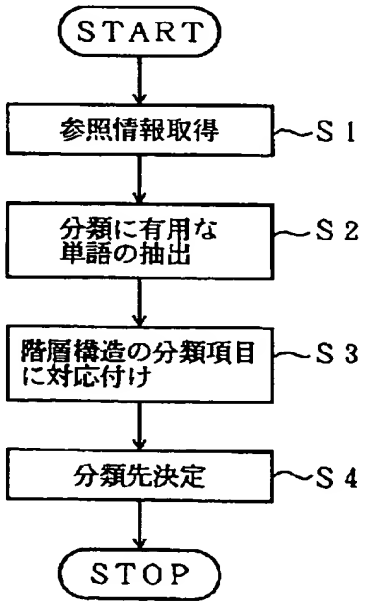
(54)【発明の名称】　情報分類方法及び装置及び情報分類プログラムを格納した記憶媒体

(57)【要約】

【課題】　分類対象の情報と当該情報が参照する情報より単語を抽出し、抽出した単語を既存の階層型知識体系に対応付けることにより、従来におけるキーワード抽出の曖昧性を起因とする分類精度の悪さ、長い学習時間による実時間処理の問題を解決した情報分類方法及び装置及び情報分類プログラムを格納した記憶媒体を提供する。

【解決手段】　本発明は、分類対象情報が参照している参照情報を取得し、分類対象情報と参照情報から分類に有用な単語を抽出し、抽出された単語を階層型知識体系の分類項目に対応付けし、対応付けされた分類項目中から分類先を決定し、分類対象情報の分類を行う。

本発明の原理を説明するための図



1

【特許請求の範囲】

【請求項１】　分類対象の情報を妥当な分類先に分類する情報分類方法において、

分類対象情報が参照している参照情報を取得し、

前記分類対象情報と前記参照情報から分類に有用な単語を抽出し、

抽出された前記単語を階層型知識体系の分類項目に対応付けし、

対応付けされた分類項目中から分類先を決定し、前記分類対象情報の分類を行うことを特徴とする情報分類方法。

【請求項２】　前記参照情報を取得する際に、

前記分類対象情報の文書を解析し、構造情報を取得し、

前記構造情報に基づいてアクセスし、リンク情報や関連情報を含む参照情報を取得する請求項１記載の情報分類方法。

【請求項３】　前記分類に有用な単語を抽出する際に、

前記分類対象情報と前記参照情報内のテキスト情報を形態素解析し、

前記形態素解析により分割された単語の品詞のうち、名詞、形容動詞を抽出し、出現頻度の大きい順にソートし、最も出現頻度の大きい単語を抽出する請求項１記載の情報分類方法。

【請求項４】　前記分類先を決定する際に、

抽出された前記単語の出現頻度と、前記階層型知識体系を用いて対応付けられた分類項目の頻度の積和を取り、最も該積和の大きいものを分類先として決定する請求項１記載の情報分類方法。

【請求項５】　分類対象の情報を妥当な分類先に分類する情報分類装置であって、

分類対象情報が参照している参照情報を取得する参照情報取得手段と、

前記分類対象情報と前記参照情報から分類に有用な単語を抽出する単語抽出手段と、

前記単語抽出手段により抽出された前記単語を階層型知識体系の分類項目に対応付けする分類項目対応付け手段と、

対応付けされた分類項目中から分類先を決定する分類先決定手段とを有することを特徴とする情報分類装置。

【請求項６】　前記参照情報取得手段は、

前記分類対象情報の文書を解析し、構造情報を取得する手段と、

前記構造情報に基づいてアクセスし、リンク情報や関連情報を含む参照情報を取得する手段を含む請求項５記載の情報分類装置。

【請求項７】　前記単語抽出手段は、

前記分類対象情報と前記参照情報内のテキスト情報を形態素解析する手段と、

前記形態素解析により分割された単語の品詞のうち、名詞、形容動詞を抽出し、出現頻度の大きい順にソートし、最も出現頻度の大きい単語を抽出する手段とを含む請求項５記載の情報分類装置。

2

【請求項８】　前記分類先決定手段は、

抽出された前記単語の出現頻度と、前記階層型知識体系を用いて対応付けられた分類項目の頻度の積和を取り、最も該積和の大きいものを分類先として決定する手段を含む請求項５記載の情報分類装置。

【請求項９】　分類対象の情報を妥当な分類先に分類する情報分類プログラムを格納した記憶媒体であって、

分類対象情報が参照している参照情報を取得する参照情報取得プロセスと、

前記分類対象情報と前記参照情報から分類に有用な単語を抽出する単語抽出プロセスと、

前記単語抽出プロセスにより抽出された前記単語を階層型知識体系の分類項目に対応付けする分類項目対応付けプロセスと、

対応付けされた分類項目中から分類先を決定する分類先決定プロセスとを有することを特徴とする情報分類プログラムを格納した記憶媒体。

【請求項１０】　前記参照情報取得プロセスは、

前記分類対象情報の文書を解析し、構造情報を取得するプロセスと、

前記構造情報に基づいてアクセスし、リンク情報や関連情報を含む参照情報を取得するプロセスを含む請求項９記載の情報分類プログラムを格納した記憶媒体。

【請求項１１】　前記単語抽出プロセスは、

前記分類対象情報と前記参照情報内のテキスト情報を形態素解析するプロセスと、

前記形態素解析により分割された単語の品詞のうち、名詞、形容動詞を抽出し、出現頻度の大きい順にソートし、最も出現頻度の大きい単語を抽出するプロセスとを含む請求項９記載の情報分類プログラムを格納した記憶媒体。

【請求項１２】　前記分類先決定プロセスは、

抽出された前記単語の出現頻度と、前記階層型知識体系を用いて対応付けられた分類項目の頻度の積和を取り、最も該積和の大きいものを分類先として決定するプロセスを含む請求項９記載の情報分類プログラムを格納した記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、情報分類方法及び装置及び情報分類プログラムを格納した記憶媒体に係り、特に、情報内の単語の頻度を分析し、当該単語を階層型知識体系に対応させることで、予め整理された分類項目の中から妥当な分類先に情報を分類する情報分類方法及び装置及び情報分類プログラムを格納した記憶媒体に関する。

【０００２】

【従来の技術】従来の情報分類技術として、当該情報内

のテキスト情報を形態素解析技術等により単語に分解し、その中から当該情報を特徴付けるような予め用意してあるキーワードを抽出し、そのキーワードに対応する分類先に分類する方法がある。この例として、電子メール整理ソフトの“Visual Mail ”の自動分類機能がある。

【０００３】また、その他の分類方法として、予め分類されている情報を答えとして特徴を学習することにより、分類する当該情報の特徴から分類先を決定する方法がある。

【０００４】

【発明が解決しようとする課題】しかしながら、上記の予め用意されているキーワードを用いて分類する方法では、当該情報から当該情報を特徴付ける妥当なキーワードを抽出するのが困難であり、分類精度が悪いという問題がある。さらに、特徴を学習することにより分類先を決定する方法は、長い学習時間が必要となり、実時間処理が必要なシステムへの応用は困難である。

【０００５】このように、上記従来の方法では、当該情報からのキーワードの曖昧性から分類精度が悪くなることが考えられる。また、予め分類されている情報の特徴の学習時間は、実時間処理の実現には問題である。本発明は、上記の点に鑑みなされたもので、分類対象の情報と当該情報が参照する情報より単語を抽出し、抽出した単語を既存の階層型知識体系に対応付けることにより、従来におけるキーワード抽出の曖昧性を起因とする分類精度の悪さ、長い学習時間による実時間処理の問題を解決した情報分類方法及び装置及び情報分類プログラムを格納した記憶媒体を提供することを目的とする。

【０００６】

【課題を解決するための手段】図１は、本発明の原理を説明するための図である。本発明（請求項１）は、分類対象の情報を妥当な分類先に分類する情報分類方法において、分類対象情報が参照している参照情報を取得し（ステップ１）、分類対象情報と参照情報から分類に有用な単語を抽出し（ステップ２）、抽出された単語を階層型知識体系の分類項目に対応付けし（ステップ３）、対応付けされた分類項目中から分類先を決定し、分類対象情報の分類を行う（ステップ４）。

【０００７】本発明（請求項２）は、参照情報を取得する際に、分類対象情報の文書を解析し、構造情報を取得し、構造情報に基づいてアクセスし、リンク情報や関連情報を含む。本発明（請求項３）は、分類に有用な単語を抽出する際に、分類対象情報と参照情報内のテキスト情報を形態素解析し、形態素解析により分割された単語の品詞のうち、名詞、形容動詞を抽出し、出現頻度の大きい順にソートし、最も出現頻度の大きい単語を抽出する。

【０００８】本発明（請求項４）は、分類先を決定する際に、抽出された単語の出現頻度と、階層型知識体系を

用いて対応付けられた分類項目の頻度の積和を取り、最も該積和の大きいものを分類先として決定する。図２は、本発明の原理構成図である。本発明（請求項５）は、分類対象の情報を妥当な分類先に分類する情報分類装置であって、分類対象情報が参照している参照情報を取得する参照情報取得手段１と、分類対象情報と参照情報から分類に有用な単語を抽出する単語抽出手段２と、単語抽出手段２により抽出された単語を階層型知識体系４の分類項目に対応付けする分類項目対応付け手段３と、対応付けされた分類項目中から分類先を決定する分類先決定手段５とを有する。

【０００９】本発明（請求項６）は、参照情報取得手段１において、分類対象情報の文書を解析し、構造情報を取得する手段と、構造情報に基づいてアクセスし、リンク情報や関連情報を含む参照情報を取得する手段を含む。本発明（請求項７）は、単語抽出手段２において、分類対象情報と参照情報内のテキスト情報を形態素解析する手段と、形態素解析により分割された単語の品詞のうち、名詞、形容動詞を抽出し、出現頻度の大きい順にソートし、最も出現頻度の大きい単語を抽出する手段とを含む。

【００１０】本発明（請求項８）は、分類先決定手段５において、抽出された単語の出現頻度と、階層型知識体系を用いて対応付けられた分類項目の頻度の積和を取り、最も該積和の大きいものを分類先として決定する手段を含む。本発明（請求項９）は、分類対象の情報を妥当な分類先に分類する情報分類プログラムを格納した記憶媒体であって、分類対象情報が参照している参照情報を取得する参照情報取得プロセスと、分類対象情報と参照情報から分類に有用な単語を抽出する単語抽出プロセスと、単語抽出プロセスにより抽出された単語を階層型知識体系の分類項目に対応付けする分類項目対応付けプロセスと、対応付けされた分類項目中から分類先を決定する分類先決定プロセスとを有する。

【００１１】本発明（請求項１０）は、参照情報取得プロセスにおいて、分類対象情報の文書を解析し、構造情報を取得するプロセスと、構造情報に基づいてアクセスし、リンク情報や関連情報を含む参照情報を取得するプロセスを含む。本発明（請求項１１）は、単語抽出プロセスにおいて、分類対象情報と参照情報内のテキスト情報を形態素解析するプロセスと、形態素解析により分割された単語の品詞のうち、名詞、形容動詞を抽出し、出現頻度の大きい順にソートし、最も出現頻度の大きい単語を抽出するプロセスとを含む。

【００１２】本発明（請求項１２）は、分類先決定プロセスにおいて、抽出された単語の出現頻度と、階層型知識体系を用いて対応付けられた分類項目の頻度の積和を取り、最も該積和の大きいものを分類先として決定するプロセスを含む。上記のように、本発明は、分類対象情報からだけではなく、当該情報が参照する参照情報から

も単語を抽出する。そのために、分類のためにより有用な単語を抽出することが可能であり、精度のよい分類を可能とする。

【００１３】また、抽出した単語を既存の階層知識体系に対応付けするため、分類対象情報から特定のキーワードが抽出されなくとも、精度のよい分類が可能となり、分類前の学習も不要となる。

【００１４】

【発明の実施の形態】図３は、本発明の分類装置の構成を示す。同図に示す分類装置は、分類対象情報が参照している参照情報を取得する参照情報取得部１、分類対象情報と参照情報から単語を抽出する単語抽出部２、単語抽出部２で抽出した単語を階層型知識体系に対応付ける検索部３、既存の階層型知識体系４、検索部３で得た分類先の候補の中から分類先を決定する分類先決定部５から構成される。

【００１５】参照情報取得部１は、入力された分類対象情報を解析して構造情報に基づいて、関連する情報、補足説明のための情報参照情報（リンク情報）を取得し、分類対象情報と当該参照情報を単語抽出部２に転送する。単語抽出部２は、取得した分類対象情報と参照情報の形態素解析を行い、単語分割と分割された単語に対して品詞を付与する。付与された品詞のうち、名詞及び形容動詞を抽出して、それらの出現頻度を求め、出現頻度の最も高い単語を検索部３に転送する。

【００１６】検索部３は、単語抽出部２で抽出された単語で階層型知識体系４を検索し、当該単語に対応する分類項目を対応付け、分類候補を取得する。分類先決定部５は、単語抽出部２で取得した単語の出現頻度と、検索部３で取得した分類項目の頻度を用いて計算を行い、その値をソートして、上位となった項目を分類先項目として決定する。

【００１７】

【実施例】以下、図面と共に本発明の実施例を説明する。以下では、インターネット上のＨＴＭＬ形式で書かれたホームページの情報を既存の階層型知識体系４として、“Yahoo Japan(http://www.yahoo.co.jp/)やＮＴＴ ＤＩＲＥＣＴＯＲＹ(http://navi.ntt.co.jp/)に代表されるインターネット上のディレクトリ型サーチエンジンを利用した場合を例として分類する過程を説明する。この場合、分類先はこのディレクトリ型サーチエンジンの各分類項目となる。

【００１８】参照情報取得部１は、分類対象情報の文書を解析してタグと呼ばれる構造情報に基づいて参照情報を取得する。図４は、本発明の一実施例のＨＴＭＬ文書の例を示す。インターネット上のホームページが同図に示すように、ＨＴＭＬ(Hyper Text Markup Language)形式と呼ばれる言語で記述されている場合、

 ～

<frame src=" URL ">

というタグに注目し、その中に記述されているＵＲＬ(Uniform Resource Locator)にアクセスすることにより、参照情報を取得する。図４の例では、

http://aaa.bbb.com/

http://ccc.ddd.com/

へアクセスし、参照情報を取得する。この参照情報はリンク情報とも呼ばれ、当該情報に関連する情報であったり、当該情報を補捉説明する情報である可能性が高い。参照情報をも考慮に入れることにより、分類対象情報に十分なテキスト情報が含まれなくとも精度の良い分類が可能になる。

【００１９】単語抽出部２では、まず、当該分類対象情報と参照情報のテキスト情報を既存技術である茶筌(http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html) のような形態素解析にかけて単語分解する。これにより、当該分類対象情報と参照情報内のテキスト情報が単語に分解され、それぞれの単語の品詞が判別される。当該分類対象情報と参照情報を単語分解した結果の例を図５に示す。この分解された単語の中から名詞、形容動詞を抽出し、出現頻度でソートし、出現頻度の大きいものを採用する。図６は、本発明の一実施例の抽出単語と出現度数の例を示す。同図の例は、抽出単語を出現頻度でソートした結果を示しており、この例では、「特許庁」という単語の出現頻度が一番大きいことが分かる。

【００２０】検索部３は、単語抽出部２で抽出した各単語に対して、階層型知識体系４の分類項目を対応付ける。具体的には、ディレクトリ検索サービスの“Yahoo Japan(http://www.yahoo.co.jp/) ”のように、単語を検索語句として入力すると、階層型知識体系４に格納されている情報の中から検索語句を含む情報とその情報が格納されている階層型知識体系４の分類項目を出力するモジュールを用いて、この検索結果から分類項目とその頻度を得る。分類項目の頻度とは検索結果の情報の中で、その分類項目に該当する情報の数を示す。

【００２１】図７は、本発明の一実施例の検索結果の例である。同図において「タイトルｎ」とあるのが、検索語句を含む情報で、「ジャンル・・・」とあるのが、情報が格納されている階層型知識体系４の分類項目である。この例で、

ジャンル：［趣味・生活］－［趣味］－［その他］－［発明］－［］－［］

に注目すると、タイトル２、７、９、１０が該当するので度数は「４」となる。このように、単語を階層型知識体系４に対応付けることにより分類先の候補を得る。

【００２２】分類先決定部では、Fw_i を単語抽出部２で得られる単語ｉの出現頻度とし、Fc_{ij}を検索部３で得られる単語ｉを検索語句とした時の分類項目ｊの頻度とした時の

【００２３】

【数１】

$$Point_j = \sum_{i \in 抽出された単語}^{7} Fw_i \times Fc_{ij}^{8}$$

【００２４】を分類項目について計算し、この値をソートし、この上位項目を採用する。図８は、本発明の一実施例の分類項目とソート結果の例を示す。この例では、［趣味・生活］－［趣味］－［その他］－［発明］が分類先として決定される。また、上記の実施例では、図３の構成要素に基づいて説明したが、この例に限定されることなく、図３の各構成要素をプログラムとして構築し、当該分類装置として利用されるコンピュータに接続されるディスク装置や、フロッピーディスクやＣＤ－ＲＯＭ等の可搬記憶媒体に格納しておき、本発明を実行する際に、インストールすることにより容易に本発明を実現することができる。

【００２５】なお、本発明は、上記の実施例に限定されることなく、特許請求の範囲内で種々変更・応用が可能である。

【００２６】

【発明の効果】上述のように、本発明によれば、分類対象情報と当該分類対象情報が参照する情報から単語を抽出し、階層型知識体系に対応付けすることにより、事前の学習をすることなしに、当該情報の分類が可能になる。さらに、特定のキーワードが抽出されなくとも精度の良い分類が可能となる。

【図面の簡単な説明】

【図１】本発明の原理を説明するための図である。

【図２】本発明の原理構成図である。

【図３】本発明の分類装置の構成図である。

【図４】本発明の一実施例のＨＴＭＬ文書の例である。

【図５】本発明の一実施例の形態素解析結果の例である。

【図６】本発明の一実施例の抽出単語と出現度数の例である。

【図７】本発明の一実施例の検索結果の例である。

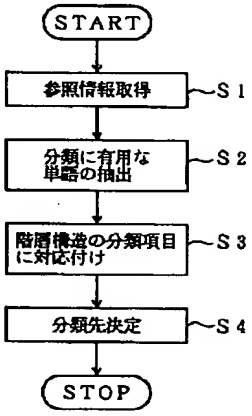
【図８】本発明の一実施例の分類項目とソート結果の例である。

【符号の説明】

1　参照情報取得部、参照情報取得手段
2　単語抽出部、単語抽出手段
3　検索部、分類項目対応付け手段
4　階層型知識体系
5　分類先決定部、分類先決定手段

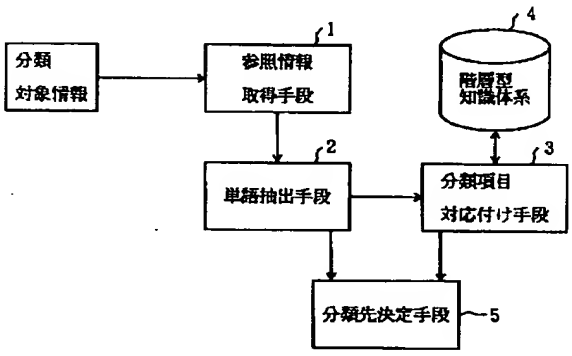
【図１】

本発明の原理を説明するための図



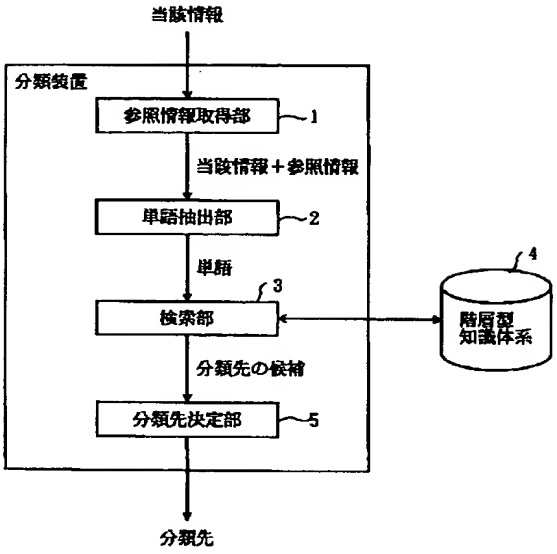
【図２】

本発明の原理構成図



【図３】

本発明の分類装置の構成図



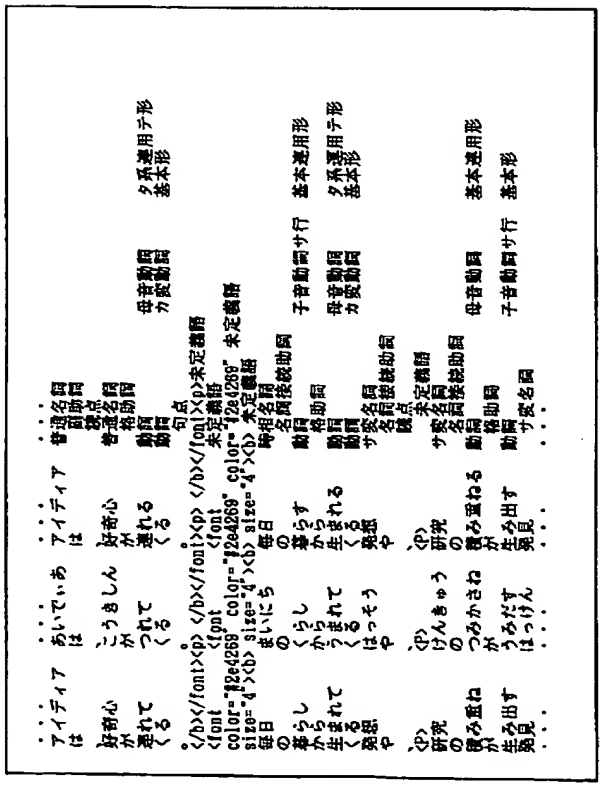
【図４】

本発明の一実施例のＨＴＭＬ文　の例

```
<HTML>
<HEAD>
   <TITLE>タイトル</TITLE>
</HEAD>
<BODY>
   <H1>題名</H1>
   <HR>
   ・・・ 本文 ・・・
   <UL>
      <LI>  <a href="http://aaa.bbb.com/">説明１</a>
      <LI>  <a href="http://ccc.ddd.com/">説明２</a>
   </UL>
   ・・・ 本文 ・・・
</BODY>
</HTML>
```

【図５】

本発明の一実施例の形態素解析結果の例



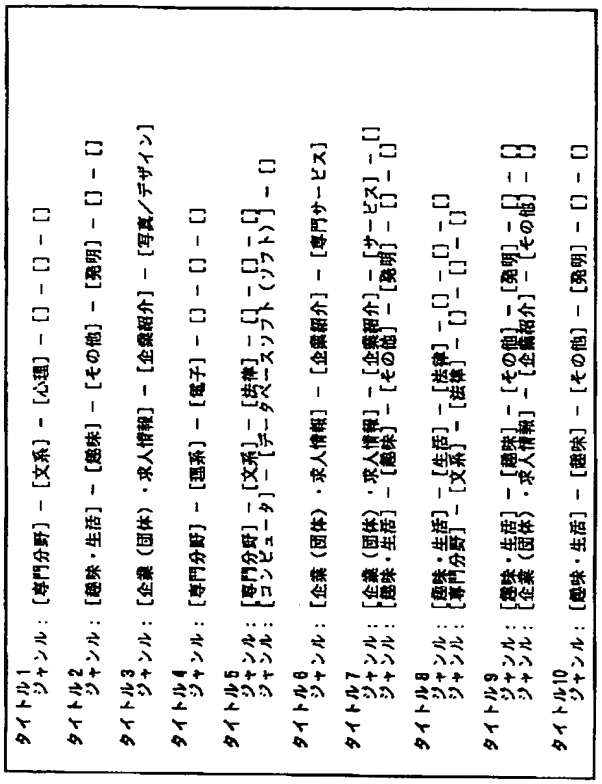
【図６】

本発明の一実施例の抽出単語と出現度数の例

単語	出現度数
特許庁	23
お知らせ	19
発表	17
官報	13
工業	12
所有権	12
種	12
掲載	11
平成	10
…	…

【図７】

本発明の一実施例の検索結果の例



【図８】

本発明の一実施例の分類項目とソート結果の例

分類項目	出現頻度の集和
［趣味・生活］－［趣味］－［その他］－［発明］	147
［専門分野］－［文系］－［法律］	77
［企業（団体）・求人情報］－［企業紹介］－［専門サービス］	58
［専門分野］－［理系］－［工学］	39
［教育・研究・文化施設］－［研究所］	32
・・・	・・・